學號：r06942098　系級：電信碩一　姓名：曾柏偉

A. PCA of colored faces

1. (.5%) 請畫出所有臉的平均。

2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

3. (.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。

4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重，請用百分比表示並四捨五入到小數點後一位。

由大排至小分別是：4.1% , 2.9%, 2.4% , 2.2%

B. Visualization of Chinese word embedding

1. (.5%) 請說明你用哪一個 word2vec 套件，並針對你有調整的參數說明那個參數的意義。
   我使用的是 gensim 的 word2vec，因為從作業四就開始使用這個套件，我覺得 size 參數的調整對於整個 word2vec 的影響是比較大的。
2. (.5%) 請在 Report 上放上你 visualization 的結果。

3. (.5%) 請討論你從 visualization 的結果觀察到什麼。
   我發現其實他的分佈其實滿有意義的，像在爸、媽附近的字詞感覺都跟親人家人有關，有些符號出現在圖片也分佈在右下角，其實經由 jieba 切出來的詞的辭意相近的在這上面的分布是距離挺接近的。

C. Image clustering

1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

	Fa-score(kaggle)	Latent dimension
Auto-encoder	73.81%	32
PCA	99.25%	300

Latent dimension is the best F1-score in my submission .

2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。

3. (.5%) visualization.npy 中前 5000 個 images 跟後 5000 個 images 來自不同 dataset。請根據這個資訊，在二維平面上視覺化 label 的分佈，接著比較和自己預測的 label 之間有何不同。

因為我的在 kaggle 上面的成績是滿高的，所以在分佈上其實並沒有差很多，我也檢查過在 predict 出來的 label 和這題資訊的 label 是全部相等的！！